Saddlepoint approximations to the mean and variance of the extended hyper geometric distribution

Eisinga, R.; Pelzer, B.

Published in:
Statistica Neerlandica

DOI:
10.1111/j.1467-9574.2010.00468.x

Statistica Neerlandica (2011) Vol. 65, nr. 1, pp. 22–31
doi:10.1111/j.1467-9574.2010.00468.x

Saddlepoint approximations to the mean and variance of the extended hypergeometric distribution

Rob Eisinga* and Ben Pelzer

Department of Social Science Research Methods, Radboud University Nijmegen, PO Box 9104, 6500 HE Nijmegen, The Netherlands

Conditional inference on 2 x 2 tables with fixed margins and unequal probabilities is based on the extended hypergeometric distribution. If the support of the distribution is large, exact calculation of the conditional mean and variance of the table entry may be computationally demanding. This paper proposes a single-saddlepoint approximation to the mean and variance. While the approximation achieves acceptable accuracy for ordinary practical purposes, an alternative saddlepoint approximation is provided that gives much closer to exact results. It improves the accuracy of current approximations to up to more than four powers of ten.

Keywords and Phrases: extended hypergeometric distribution, saddlepoint approximation, 2 x 2 table.

1 Introduction

The conditional approach to inference for 2×2 contingency tables with fixed marginal totals and unequal probability parameters is based on the extended hypergeometric distribution (Harkness, 1965; Johnson, Kemp and Kotz, 2005). This distribution, also known as the Fisher non-central hypergeometric distribution (Fog, 2008), has found wide application in biostatistical, epidemiological, and social research. It gave rise to Fisher's exact test (Fisher, 1935) and it is used, for example, in power calculation and sample size determination (Munoz and Rosner, 1984), the estimation of false discovery rates (Tsai, Hsueh and Chen, 2003), and the analysis of ecological data (Wakefield, 2004; Xu *et al.*, 2008). Exact calculation of the conditional mean and variance of an entry of the table, given the marginal totals, requires complete enumeration of all possible cell frequencies consistent with the observed table margins. If the number of admissible tables is large, exact calculation of the mean and variance becomes computationally burdensome or even infeasible, especially if expected values of the moments are required, computed over the distribution of a margin.

*r.eisinga@maw.ru.nl

This paper uses a single-saddlepoint approximation to the distribution of two independent binomials, conditioned on fixed sum, to obtain approximations to the mean and variance of an extended hypergeometric random variable. The derivation of the approximation is quite direct, and the result is shown to be algebraically equivalent to previously reported approximations to the mean. The paper subsequently proposes an alternative saddlepoint approximation based on an improvement, not addressed before, to the first-order variance approximation. The numerical evidence indicates that this novel approach substantially reduces the error of approximation and provides amazingly accurate results, particularly when it comes to approximating the mean.

The paper is organized as follows. Section 2 considers relevant properties of the extended hypergeometric distribution and reviews current approximations to the mean and variance. Section 3 discusses a single-saddlepoint approximation to the moments. The improved saddlepoint method is presented in section 4 and section 5 presents results of a numerical investigation. Conclusions are in section 6.

2 Extended hypergeometric distribution

Consider two independent binomial random variables Y_j with fixed denominators n_j and probabilities π_j, that is $Y_j \sim \text{Binomial}(n_j, \pi_j), j = 0, 1$. Let m be the sum of the observed values of Y_0 and Y_1. Conditional inference is based on the distribution of Y_0 given that $M = Y_0 + Y_1 = m$. This distribution is an extended hypergeometric distribution (Harkness, 1965; Plackett, 1981; McCullagh and Nelder, 1992; Fog, 2008), wherein the probability mass function $p(Y_0 = y_0 \mid M = m)$ depends on π_j only through the odds ratio parameter $\psi = \pi_0(1-\pi_1)/\{\pi_1(1-\pi_0)\}$,

$$
\begin{aligned}
p(Y_0 = y_0 \mid M = m) &= \frac{p(Y_0 = y_0)\, p(Y_1 = m - y_0)}{\sum_{i=l}^{s} p(Y_0 = i)\, p(Y_1 = m - i)} \\
&= \frac{\binom{n_0}{y_0} \pi_0^{y_0} (1-\pi_0)^{n_0 - y_0} \binom{n_1}{m - y_0} \pi_1^{m - y_0} (1-\pi_1)^{n_1 - m + y_0}}{\sum_{i=l}^{s} \binom{n_0}{i} \pi_0^{i} (1-\pi_0)^{n_0 - i} \binom{n_1}{m - i} \pi_1^{m - i} (1-\pi_1)^{n_1 - m + i}} \\
&= \frac{\binom{n_0}{y_0} \binom{n_1}{m - y_0} \psi^{y_0}}{\sum_{i=l}^{s} \binom{n_0}{i} \binom{n_1}{m - i} \psi^{i}},
\end{aligned}
$$

where i indicates summation over all permissible values of y_0, with lower bound $l = \max(0, m - n_1)$ and upper bound $s = \min(n_0, m)$. The conditional expectation of Y_0, given $M = m$, can be expressed as

$$
\mu = E(Y_0) = \frac{\sum_{i=l}^{s} i p(Y_0 = i) p(Y_1 = m - i)}{p(M = m)},
$$

and the conditional variance of Y_0, given $M = m$, as

$$\sigma^2 = \mathrm{var}(Y_0) = \frac{\sum_{i=l}^{s} i^2 p(Y_0 = i) p(Y_1 = m - i)}{p(M = m)} - \left\{ \frac{\sum_{i=l}^{s} i\, p(Y_0 = i) p(Y_1 = m - i)}{p(M = m)} \right\}^2. \quad (1)$$

As these expressions involve summations over $s - l + 1$ terms, the mean and variance are awkward to compute if the number of terms is large. To save computation time, recursive algorithms have been proposed for calculating the mean and the variance (Satten and Kuppen, 1990; Liao, 1992), and for computing the distribution of Y_0 and sampling from it (Liao and Rosen, 2001; Fog, 2008). Calculating the moments can still be time consuming if the support of the distribution has a large number of elements and precise calculation is required, even with special algorithms and modern computing power (Agresti, 2002). This is particularly true when dealing with nested summations, as in the computation of expected values of the moments. The reference set of tables on which the conditional distribution is defined may then become enormous. For instance, Wakefield (2004) analysed sixty-four 2×2 tables with grand totals ranging from 4,421 to 217,967. In this case, the number of possible tables to be processed in computing the marginal expectation of the variance is in the order of 31.7×10^9. Highly accurate approximations that avoid exact computation are practically very useful in such case.

Approximations to the mean and variance of the extended hypergeometric distribution have been thoroughly investigated in the biostatistical literature (Stevens, 1951; Cornfield, 1956; Levin, 1984; McCullagh, 1984; Gart, 1987; McCullagh and Nelder, 1992). The results show that a first-order asymptotic approximation to μ, $\tilde{\mu}$, can be found as the appropriate solution of the quadratic equation

$$\psi = \frac{\tilde{\mu}(\tilde{\mu} + n_1 - m)}{(m - \tilde{\mu})(n_0 - \tilde{\mu})}.$$

Explicitly, when $\psi \neq 1$, $\tilde{\mu}$ is the solution to

$$\tilde{\mu} = \frac{1}{2} \left\{ b - \left(b^2 - 4ac \right)^{1/2} \right\} a^{-1}, \quad (2)$$

where $a = \psi - 1, b = (\psi - 1)(n_0 + m) + n$, and $c = \psi n_0 m$. For $\psi = 1$, $\tilde{\mu} = \mu = n_0 m / n$, the exact expected value of the central hypergeometric distribution. Harkness (1965) was the first to note the exact relationship between μ and σ^2. He expressed the mean in terms of the variance as

$$\mu = \frac{1}{2} \left\{ b - \left(b^2 - 4ac - 4a^2\sigma^2 \right)^{1/2} \right\} a^{-1}.$$

This relationship was exploited by Levin (1984) to obtain a correction to $\tilde{\mu}$ by re-expressing the solution for μ in terms of σ^2 in the form

$$\mu = \tilde{\mu} + \frac{1}{2} d \left\{ 1 - \left(1 - 4a^2\sigma^2 d^{-2} \right)^{1/2} \right\} a^{-1}, \quad (3)$$

where $d = (b^2 - 4ac)^{1/2}$. Hence $\tilde{\mu} < \mu$. A first-order asymptotic approximation to σ^2 is

$$\tilde{v} = \left(\frac{n}{n-1} \right) \left(\frac{1}{\tilde{\mu}} + \frac{1}{n_0 - \tilde{\mu}} + \frac{1}{m - \tilde{\mu}} + \frac{1}{n_1 - m + \tilde{\mu}} \right)^{-1},$$

where, following McCullagh (1984), the correction factor $n/(n-1)$ is included to ensure that $\tilde{v}=\sigma^2$ if $\psi=1$. The improvement to $\tilde{\mu}$ suggested by Levin (1984) is to replace σ^2 in Equation 3 by the approximation to σ^2, $\tilde{v}$, so that the resulting $\tilde{\mu}'$ corrects the underestimation of μ by $\tilde{\mu}$.

Levin (1984) also obtained the first-order approximation to Equation 3, with σ^2 replaced by $\tilde{v}$,

$$\tilde{\mu}''=\tilde{\mu}+a\tilde{v}d^{-1},$$

and in a subsequent paper Levin (1990) pointed out that this approximation derives from a double-saddlepoint approximation to the conditional score function. He argued that if the score function is of order n, and the double-saddlepoint correction is of order 1, the error in the approximation is of order n^{-1}, and he mentioned that this solves a mystery that puzzled Gart (1987) as to why this correction is so accurate.

3 Single-saddlepoint approximation

Saddlepoint methods, first laid out in the pioneering paper of Daniels (1954), have become popular for approximating probability density functions and tail probabilities (Butler, 2007). With respect to the current issue, there are several ways to obtain a saddlepoint approximation to the mean and variance of Y_0, given $M=m$. The moment generating function of the hypergeometric distribution is a ratio of Gauss hypergeometric functions (Johnson *et al.*, 2005) and these ${}_2F_1$ can readily be approximated by the saddlepoint method (Butler and Wood, 2002; Butler, 2007). Another option is to employ a double-saddlepoint approximation, one for the joint distribution and another for the marginal (Davison, 1988; Levin 1990; Butler, 2007).

We take a simpler and straightforward approach and seek an approximation to the probability mass function of the sum of the two independent binomials $M=Y_0+Y_1$, denoted $p(M=m)$. The cumulant-generating function of this convolution is

$$K(u)=n_0\ln\{1-\pi_0+\pi_0\exp(u)\}+n_1\ln\{1-\pi_1+\pi_1\exp(u)\}.$$

Let $q_j=\pi_j\exp(u)/\{1-\pi_j+\pi_j\exp(u)\}$, $j=0,1$. The first-order saddlepoint approximation to the mass function of M is then given by

$$\tilde{p}(M=m)=\{2\pi K''(\tilde{u})\}^{-1/2}\exp\{K(\tilde{u})-\tilde{u}m\},$$

where $\tilde{u}=\tilde{u}(m)$, the saddlepoint, is the unique value of u satisfying the saddlepoint equation $K'(u)=m$, with $K'(u)=n_0q_0+n_1q_1$ and $K''(u)=n_0q_0(1-q_0)+n_1q_1(1-q_1)$ being the first- and second-order derivatives of $K(u)$ with respect to u.

As the second cumulant $K''(u)$ is the variance of Y_j, $K''(u)>0$ for all u. Hence $K(u)$ is a convex function, and this implies that the equation $K'(u)=m$ has at most one solution. Consequently, the approximate $\tilde{p}(M=m)$ may be obtained explicitly as

$$\tilde{p}(M=m)=\{2\pi K''(\tilde{u})\}^{-1/2}\{\exp(\tilde{u})\}^{-m}\{1-\pi_0+\pi_0\exp(\tilde{u})\}^{n_0}\{1-\pi_1+\pi_1\exp(\tilde{u})\}^{n_1},$$

where

$$\tilde{u}=\ln\left\{\frac{1}{2}\left(-b'+\left(b'^2-4a'c'\right)^{1/2}\right)a'^{-1}\right\},$$

is determined by solving $K'(u)=m$ for u, with $a'=(n-m)\pi_0\pi_1, b'=m\pi_0\pi_1-(n-m)\pi_0\pi_1+n_0\pi_0+n_1\pi_1-m(\pi_0+\pi_1)$, and $c'=-m(1-\pi_0)(1-\pi_1)$. As Y_0 and Y_1 are independent random variables, the saddlepoint approximation to the conditional mean and variance of Y_0 are obtained as

$$\tilde{\mu}=n_0q_0$$

and

$$\begin{aligned}\tilde{v}&=\left(\frac{n}{n-1}\right)\left(\frac{1}{n_0q_0(1-q_0)}+\frac{1}{n_1q_1(1-q_1)}\right)^{-1}\\&=\left(\frac{n}{n-1}\right)\left(\frac{1}{\tilde{\mu}}+\frac{1}{n_0-\tilde{\mu}}+\frac{1}{m-\tilde{\mu}}+\frac{1}{n_1-m+\tilde{\mu}}\right)^{-1},\end{aligned}$$

respectively. To obtain an expression in terms of the odds ratio ψ, the single-saddlepoint approximation to the mean may be re-written as

$$\tilde{\mu}=\frac{1}{2}\left\{b-\left(b^2-4ac\right)^{1/2}\right\}a^{-1}$$

with the constants defined as in Equation 2. Hence the approximate mean discussed above is equivalent to the approximate mean that results from a single-saddlepoint approximation to the mass function of the convolution of two independent binomials. Also, there is an exact parametric relation between the mean and variance and the saddlepoint approximation to the mean, equivalent to Equation 3. If, following Levin (1984), in this relationship σ^2 is replaced by $\tilde{v}$, we obtain an improved approximate mean

$$\tilde{\mu}'=\tilde{\mu}+\frac{1}{2}d\left\{1-\left(1-4a^2\tilde{v}d^{-2}\right)^{1/2}\right\}a^{-1},$$

where $d=(b^2-4ac)^{1/2}$.

4 Alternative saddlepoint approximation

The exact conditional variance given in Equation 1 can be re-written (see Appendix) so as to yield

$$\begin{aligned}\sigma^2=&\frac{n_0\pi_0p(M^{(-1)}=m-1)+n_0\pi_0(n_0-1)\pi_0\,p(M^{(-2)}=m-2)}{p(M=m)}\\&-\left\{\frac{n_0\pi_0\,p(M^{(-1)}=m-1)}{p(M=m)}\right\}^2,\end{aligned}\tag{4}$$

where $M^{(-1)} = Y_0^{(-1)} + Y_1$, $M^{(-2)} = Y_0^{(-2)} + Y_1$, with $Y_0^{(-1)} \sim \text{Binomial}(n_0 - 1, \pi_0)$ and $Y_0^{(-2)} \sim \text{Binomial}(n_0 - 2, \pi_0)$. This expression suggests an alternative approximation to the variance based on a single-saddlepoint approximation to each of the three mass functions of M, $M-1$, and $M-2$

$$\tilde{v}' = \frac{n_0\pi_0\tilde{p}(M^{(-1)} = m-1) + n_0\pi_0(n_0-1)\pi_0\tilde{p}(M^{(-2)} = m-2)}{\tilde{p}(M=m)} - \left\{\frac{n_0\pi_0\tilde{p}(M^{(-1)} = m-1)}{\tilde{p}(M=m)}\right\}^2.$$

This approximation may subsequently be exploited to adapt the approximate mean using

$$\tilde{\mu}''' = \tilde{\mu}' + \frac{1}{2}d\left\{1 - \left(1 - 4a^2\tilde{v}'d^{-2}\right)^{1/2}\right\}a^{-1}.$$

There are at least two strategies to minimize the error of the saddlepoint approximation to the probability mass function of M. For a single-saddlepoint approximation, the error is of order $O(n^{-1})$,

$$p(M=m) = \tilde{p}(M=m)\{1 + O(n^{-1})\}.$$

Unlike the normal approximation, whose error is absolute and of order $O(n^{-1/2})$, the error of the saddlepoint approximation is relative, which implies that the approximation improves as the margins of the table increase. The order of accuracy of $\tilde{p}(M=m)$ can be improved by normalizing it to sum to unity. Daniels (1954) shows that a normalized saddlepoint approximation to the mass function of M will be of the order $O(n^{-3/2})$. Alternatively, a higher order and generally more accurate approximation can be obtained by including adjustments for the third and fourth cumulants (Daniels, 1987; Akahira and Takahashi, 2001). The second-order saddlepoint approximation uses the correction term

$$\tilde{\tilde{p}}(M=m) = \tilde{p}(M=m)\left\{1 + \frac{1}{8}\frac{K''''(\tilde{u})}{\{K''(\tilde{u})\}^2} - \frac{5}{24}\frac{\{K'''(\tilde{u})\}^2}{\{K''(\tilde{u})\}^3} + O\left(n^{-2}\right)\right\},$$

where

$$K'''(u) = n_0q_0(1-q_0)(1-2q_0) + n_1q_1(1-q_1)(1-2q_1)$$

and

$$K''''(u) = n_0q_0(1-q_0)\{1-6q_0(1-q_0)\} + n_1q_1(1-q_1)\{1-6q_1(1-q_1)\}$$

with $q_j = \pi_j \exp(u)/\{1 - \pi_j + \pi_j \exp(u)\}$, $j = 0, 1$. The same correction is used to enhance the accuracy of the mass function approximations to $M-1$ and $M-2$, and thereby to improve the approximate mean and variance.

Table 1. Saddlepoint approximations, error rate (ε), and exact mean and variance

	ψ	$n_0=10, m=10, n=20$		$n_0=100, m=100, n=200$		$n_0=1000, m=1000, n=2000$	
			ε		ε		ε
$\tilde{\mu}'$	2.0	5.90308	$0.0^5 2237$	58.62175910	$0.0^8 1744$	585.82935295421	$0.0^{11} 1703$
$\tilde{\mu}''$		5.90301	$0.0^4 1449$	58.62175252	$0.0^6 1139$	585.82935230305	$0.0^8 1113$
$\tilde{\mu}'''$		5.90309	$0.0^6 5773$	58.62175919	$0.0^{10} 3463$	585.82935295521	$0.0^{14} 5822$
μ		5.90310		58.62175920		585.82935295520	
$\tilde{v}$		1.27705	$0.0^3 2914$	12.1929993	$0.0^5 2370$	121.38103408	$0.0^7 2326$
$\tilde{v}'$		1.27733	$0.0^4 7522$	12.1930276	$0.0^7 4707$	121.38103691	$0.0^{10} 6330$
σ^2		1.27742		12.1930282		121.38103690	
$\tilde{\mu}'$	6.0	7.21287	$0.0^4 3999$	71.11589808	$0.0^7 2812$	710.20716629785	$0.0^{10} 2728$
$\tilde{\mu}''$		7.21160	$0.0^3 2170$	71.11578406	$0.0^5 1631$	710.20715502088	$0.0^7 1591$
$\tilde{\mu}'''$		7.21312	$0.0^5 6302$	71.11590005	$0.0^9 3291$	710.20716631725	$0.0^{13} 3170$
μ		7.21316		71.11590007		710.20716631723	
$\tilde{v}$		1.0834	$0.0^2 2543$	10.344579	$0.0^4 1890$	102.98005402	$0.0^6 1843$
$\tilde{v}'$		1.0857	$0.0^3 4007$	10.344773	$0.0^6 2212$	102.98007302	$0.0^9 2057$
σ^2		1.0862		10.344775		102.98007300	
		$n_0=50, m=100, n=200$		$n_0=150, m=50, n=200$		$n_0=100, m=10, n=200$	
$\tilde{\mu}'$	2.0	31.417051528	$0.0^6 3618$	41.90374962	$0.0^6 9375$	6.5990427	$0.0^5 1521$
$\tilde{\mu}''$		31.417054900	$0.0^6 4692$	41.90375060	$0.0^6 9609$	6.5990428	$0.0^5 1548$
$\tilde{\mu}'''$		31.417040160	$0.0^{11} 8922$	41.90371032	$0.0^9 3486$	6.5990322	$0.0^7 6623$
μ		31.417040161		41.90371033		6.5990326	
$\tilde{v}$		8.9378765	$0.0^3 4738$	5.575927	$0.0^2 2289$	2.15592	$0.0^2 1409$
$\tilde{v}'$		8.9336437	$0.0^8 9010$	5.563191	$0.0^6 8302$	2.15275	$0.0^4 6026$
σ^2		8.9336438		5.563195		2.15288	

Note: $\tilde{\mu}'$ and $\tilde{v}$ are single-saddlepoint approximations, $\tilde{\mu}''$ is the double-saddlepoint approximation by Levin (1990), and $\tilde{\mu}'''$ and $\tilde{v}'$ are obtained using three single-saddlepoint approximations as described in section 4. The error rate is $\varepsilon = |\omega - \omega_{\text{approx}}| / |\omega|$.

5 Accuracy assessment

The accuracy of the single-saddlepoint approximations $\tilde{\mu}'$ and $\tilde{v}$, the double-saddlepoint approximation given by Levin (1990), and the approximations $\tilde{\mu}'''$ and $\tilde{v}'$ proposed here, based on three single-saddlepoint approximations, was examined numerically for a wide variety of marginal totals and odds ratios in the range $10 \leq n_0, m, n \leq 2000$ and $1 < \psi \leq 80$. Table 1 presents the results for 2×2 tables with equal marginal totals $n_0 = m = \frac{1}{2}n = 10, 100, 1000$ and $\psi = 2, 6,$ as well as for tables with unequal totals and rather imbalanced data, $\psi = 2$.

As can be seen, the saddlepoint methods are shown to result in highly accurate approximations to the exact moments, especially the mean, even for tables with relatively small marginal totals such as those reported here. The numerical evidence also suggests that both the single-saddlepoint approximation $\tilde{\mu}'$ and the double-saddlepoint approximation $\tilde{\mu}''$ do not approximate the exact results nearly as well as $\tilde{\mu}'''$, which uses three single-saddlepoint approximations. The latter raise the accuracy of the approximation to up to more than four powers of ten. Similar results go for the approximate variance. According to the error rates in Table 1, the approximation $\tilde{v}'$ is a major improvement to the first-order variance approximation $\tilde{v}$. The improvements

$\tilde{\mu}'''$ and $\tilde{v}'$ achieved tend to increase as the table margins increase in size and they provide much closer to exact results for tables with equal and unequal marginal totals as well as for tables with imbalanced data.

The same conclusion is obtained for the almost complete range of marginal totals ($10 \leq n_0, m, n \leq 2000$) and odds ratios ($1 < \psi \leq 80$). In 99.8% of the 55.7×10^4 tables analysed, $\tilde{\mu}'''$ had less relative error than both $\tilde{\mu}'$ and $\tilde{\mu}''$, and $\tilde{v}'$ less error than $\tilde{v}$. The 0.2% exception concerns, for the main part, highly skewed tables with $m \ll n$, $m = n_1$, and ψ large.

6 Conclusion

This paper considers saddlepoint approximations to the mean and variance of the extended hypergeometric distribution. It shows that the approximate mean discussed in the biostatistical literature equals the approximate mean resulting from a single-saddlepoint approximation to the convolution of two independent binomials. A novel approximation is provided based on three single-saddlepoint approximations. This alternative method substantially improves the accuracy of the approximate mean and variance and offers highly accurate results.

The presented approximation is a particularly useful tool when dealing with nested summations and accuracy and speed are needed. It may be used to accelerate an EM-type algorithm by reducing the time spend in the E-step, which depends on the number of admissible tables. Finally, returning to the example mentioned above, computing the expectation of the exact conditional variance for the data reported by Wakefield (2004) is very demanding, and can be a matter of days. The computation time required to obtain the saddlepoint approximations is essentially negligible as compared with exact calculation. Replacing the inner summation with a saddlepoint approximation produces virtually identical results, within seconds.

Acknowledgements

The authors are grateful to Bruce Levin and Agner Fog for comments on a previous version of this manuscript and suggestions for improvements.

Appendix

Derivation of Equation 4

The derivation of the second term on the RHS of Equation 1 is reported in Tsai *et al.* (2003). Since $i\binom{n_0}{i} = n_0\binom{n_0-1}{i-1}$, the first term on the RHS of Equation 1 can be expressed as

$$\frac{\sum_{i=l}^{s} i^2 p(Y_0=i)p(Y_1=m-i)}{p(M=m)} = \frac{n_0\pi_0 \sum_{i=l}^{s} ip(Y_0^{(-1)}=i-1)p(Y_1=m-i)}{p(M=m)}$$
$$= \frac{\begin{array}{c}(n_0\pi_0 \sum_{i=l}^{s} p(Y_0^{(-1)}=i-1)p(Y_1=m-i) \\ + n_0\pi_0(n_0-1)\pi_0 \sum_{i=l}^{s} p(Y_0^{(-2)}=i-2)p(Y_1=m-i))\end{array}}{p(M=m)}$$
$$= \frac{n_0\pi_0 p(M^{(-1)}=m-1) + n_0\pi_0(n_0-1)\pi_0\, p(M^{(-2)}=m-2)}{p(M=m)},$$

where $M^{(-1)} = Y_0^{(-1)} + Y_1, \quad M^{(-2)} = Y_0^{(-2)} + Y_1$, with $Y_0^{(-1)} \sim \text{Binomial}(n_0-1, \pi_0)$ and $Y_0^{(-2)} \sim \text{Binomial}(n_0-2, \pi_0)$. Note that this derivation includes the derivation of the second term on the RHS of Equation 1. This proofs Equation 4.

References

Agresti, A. (2002), Exact inference for categorical data: recent advances and continuing controversies, *Statistics in Medicine* **20**, 2709–2722.

Akahira, M. and K. Takahashi (2001), A higher order large-deviation approximation for the discrete distributions, *Journal of the Japan Statistical Society* **31**, 257–267.

Butler, R. W. (2007), *Saddlepoint approximations with applications*, Cambridge University Press, Cambridge, MA.

Butler, R. W. and A. T. A. Wood (2002), Laplace approximations for hypergeometric functions with matrix argument, *Annals of Statistics* **30**, 1155–1177.

Cornfield, J. (1956), A statistical problem arising from retrospective studies, *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* **4**, 134–148.

Daniels, H. E. (1954), Saddlepoint approximations in statistics, *Annals of Mathematical Statistics* **25**, 631–650.

Daniels, H. E. (1987), Tail probability approximations, *International Statistical Review* **55**, 37–48.

Davison, A. C. (1988), Approximate conditional inference in generalized linear models, *Journal of the Royal Statistical Society. Series B* **50**, 445–461.

Fisher, R. A. (1935), The logic of inductive inference, *Journal of the Royal Statistical Society* **98**, 35–54.

Fog, A. (2008), Sampling methods for Wallenius' and Fisher's noncentral hypergeometric distributions, *Communications in Statistics – Simulation and Computation* **37**, 241–257.

Gart, J. J. (1987), The equivalence of two corrections to the approximate mean of an entry in a contingency table, *Biometrika* **74**, 661–663.

Harkness, W. L. (1965), Properties of the extended hypergeometric distribution, *Annals of Mathematical Statistics* **36**, 938–945.

Johnson, N. L., A. W. Kemp and S. Kotz (2005), *Univariate discrete distributions*, Wiley, Hoboken NJ.

Levin, B. (1984), Simple improvements on Cornfield's approximation to the mean of a noncentral hypergeometric random variable, *Biometrika* **71**, 630–632.

Levin, B. (1990), The saddlepoint correction in conditional logistic likelihood analysis, *Biometrika* **77**, 275–285.

Liao, J. (1992), An algorithm for the mean and variance of the noncentral hypergeometric distribution, *Biometrics* **48**, 889–892.

Liao, J. G. and O. Rosen (2001), Fast and stable algorithms for computing and sampling from the noncentral hypergeometric distribution, *American Statistician* **55**, 366–369.

McCullagh, P. (1984), On the elimination of nuisance parameters in the proportional odds model, *Journal of the Royal Statistical Society. Series B* **46**, 250–256.

McCullagh, P. and J. A. Nelder (1992), *Generalized linear models* (2nd edn), Chapman and Hall, London.

Munoz, A. and B., Rosner (1984), Power and sample size for a collection of 2×2 tables, *Biometrics* **40**, 995–1004.

Plackett, R. L. (1981). *The analysis of categorical data*, 2nd edn, Griffin, London.

Satten, G. A. and L. L. Kuppen (1990), Continued fraction representation for expected cell counts of a 2×2 table: a rapid and exact method for conditional maximum likelihood estimation, *Biometrics* **46**, 217–223.

Stevens, W. L. (1951), Mean and variance of an entry in a contingency table, *Biometrika* **38**, 468–470.

Tsai, C-A., H-m. Hsueh and J. J. Chen (2003), Estimation of false discovery rates in multiple testing: application to gene microarray data, *Biometrics* **59**, 1071–1081.

Wakefield, J. (2004), Ecological inference for 2×2 tables (with discussion), *Journal of the Royal Statistical Society. Series A* **167**, 385–445.

Xu, J., Y. Yang, Z. Ying and J. Ott (2008), Testing linkage disequilibrium from pooled DNA: a contingency table perspective, *Statistics in Medicine* **27**, 5801–5815.

Received: December 2009. Revised: May 2010.